3D Enriched Fragment Library Design for X-ray Crystallography

Michael P. Mazanetz, Paul A McEwan, John Barker, Michael Bodkin

Evotec (UK) Ltd., 114 Innovation Drive, Milton Park, Abingdon, Oxfordshire OX14 4RZ, United Kingdom

Abstract

This poster describes a cheminformatics approach using KNIME [1] to design a screening library which represents diversity in fragment chemistry [2] and geometry as well as considers chemical features amenable to X-ray crystallography. This novel approach searches across commercially available fragment-like chemical space to identify diverse compounds in terms of chemical features, conformational space and molecular features pertaining to 3D-shape. Library diversity is enhanced through the measure of our 'Cocktail Score' as the algorithm optimally searches commercial chemical space to finally produce a library which is enriched with 3D-like structures determined by thresholds in the sum of the normalised principal moments of inertia (NPR1+NPR2) and the plane of best-fit (PBF).

Defining Fragment Space for X-ray Screening

There are a number of considerations that need to be taken into account in the generation of a compound library for fragment screening. The Astex 'Rule of Three' [3] is a useful guide to describe fragment properties and include; molecular weight ≤300 Da, clogP ≤3 and H-bond donors ≤3. Other considerations include the removal of undesirable functionality which can be defined using the Quantitative Estimate of Druglikeness (QED) metric, the size and diversity of the library, focussed sets, particular strategies for subsequent fragment evolution and optimisation. There are other molecular attributes and physicochemical properties which need to be tailored to be suitable for screening by X-ray crystallography. The fragments need to have sufficient aqueous solubility such that they can be effectively screened at high concentration.

The size of the library for X-ray screening is limited to only relatively low numbers of fragments (typically less than 1,000), even with automation and fragment pooling (called cocktailing) due to the nature of the X-ray method. For compounds to be screened as cocktails in crystal the presence of bromine and sulphur atoms can be of benefit due to unique its dispersion signal and 3D shape diversity for the fragments in each cocktail can aid the direct deconvolution of hits.

The fragment attributes amenable for X-ray screening are summarised in Table 1. These criteria can then be used to filter commercial chemical libraries to define a potential pool of compounds to select from to build a fragment library. We chose 17 commercial chemical suppliers, which resulted in over 544,000 compounds to consider for our fragment library.

Attribute	Value	Attribute	Value
QED	≥ 0.6	Number of hydrogen-bond donors	0-10
cLogS	> -4.0	Number of hydrogen-bond acceptors	0-10
Molecular weight	150-350	Number formal charges	0-4
Number of heavy atoms	5-19	Sum of formal charges	-2 – 2
Number of rotatable bonds	0-12	Allowed Elements	H, C, N, O, F, S, Cl, Br, I
Number of connected unbranched non-ring atoms	0-8	Dis-allowed Metals	Sc,Ti,V,Cr,Mn,Fe,Co,Ni, Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru, Rh,Pd,Ag,Cd

Table 1: The molecular attributes and physicochemical properties used as a filter to identify potential compounds for the X-ray fragment library.

Describing 3D Space

3D character has been proposed as an additional criteria for fragment libraries and may be assessed by the fraction of sp^3 carbon atoms (Fsp^3), the normalised principal moments of inertia (NPR1 and 2) [4] the plane of best fit (PBF) [5]. There are limitations to these methods, particularly Fsp^3 which is too crude a measure for 3D-likeness, Fig. 1. NPR1 and NPR2 are calculated from the principal moments of inertia of a low energy conformation. Values can only lie within a special triangular area, and 3D-lkeness can be measured as the sum of NPR1 and NPR2, Fig. 2. This metric too has its challenges, for example, the 3D-likeness of compounds which have a summed valued close to 1 could either be rod-like or spherical, Fig. 2.

PBF which quantifies the average distance in angstroms of all heavy atoms away from the plane of best fit through all heavy atoms (Fig. 3) does not correlate with Fsp^3 which is more a measure of complexity.

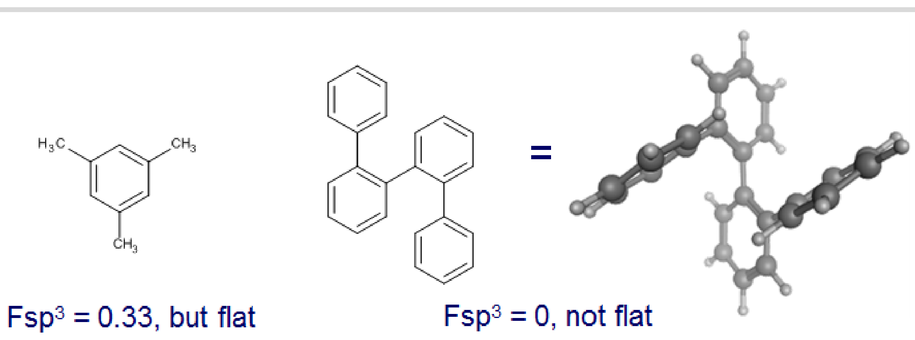

Figure 1: Examples showing the variability in the Fsp3 metric

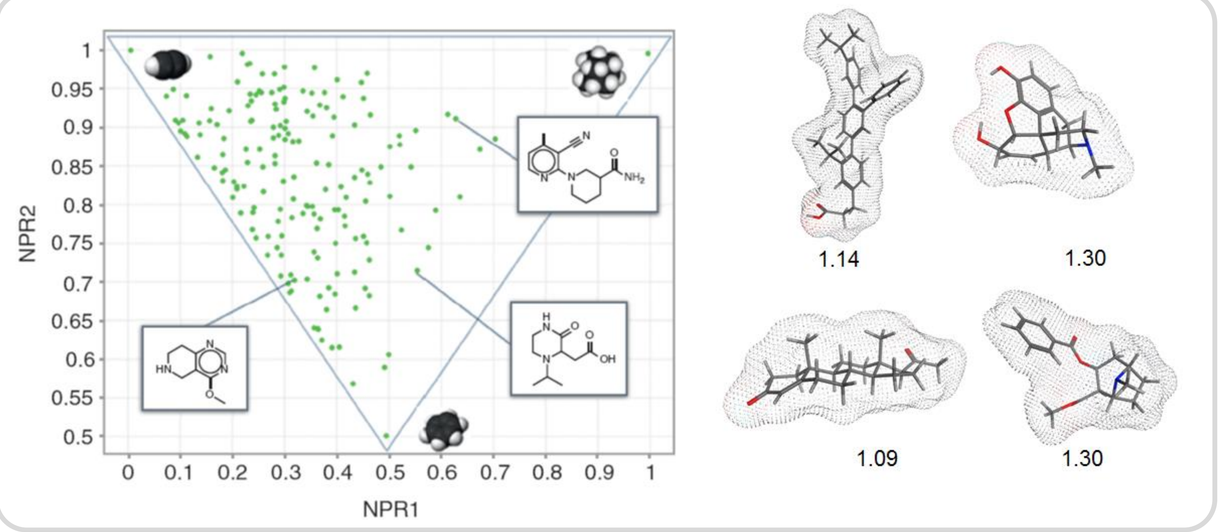

Figure 2: *(left)* The chart shows an arbitrary collection of compounds and whether a molecule is near to stick- (top left), disc- (bottom), or sphere-like shape (top right) "3D-like" compounds should lie closer to the top right corner of the triangle: sum of NPR1+NPR2 =1.0 along the left side of the triangle, 2.0 in the top right corner). *(right)* 3D representations of 4 compounds and their associated sum of NPR1+NPR2.

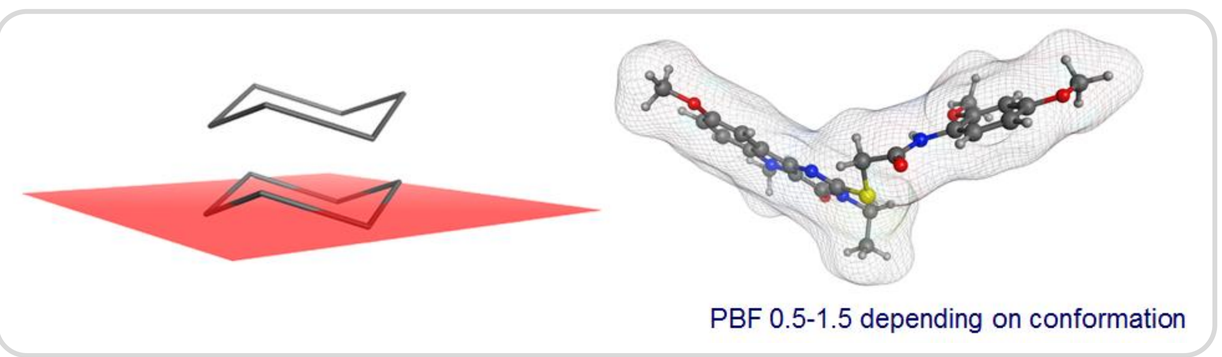

Figure 3: *(left)* Plane of Best Fit, showing how a plane (red) can be placed through a cyclohexane conformer from which the atom distances to the plane can be measured. *(right)* Illustration of the conformational dependency of the PBF score.

Our approach was to generate a conformational ensemble for each compound in the 544,000 pool and then to determine an average value for the 3D-like features. Compounds filtered from the vendor library could be grouped into a 3D-like class if the PBF score was >= 0.6 and the sum of NPR1+2 was >= 1.07 and all other compounds classed as 2D-like, Fig. 4.

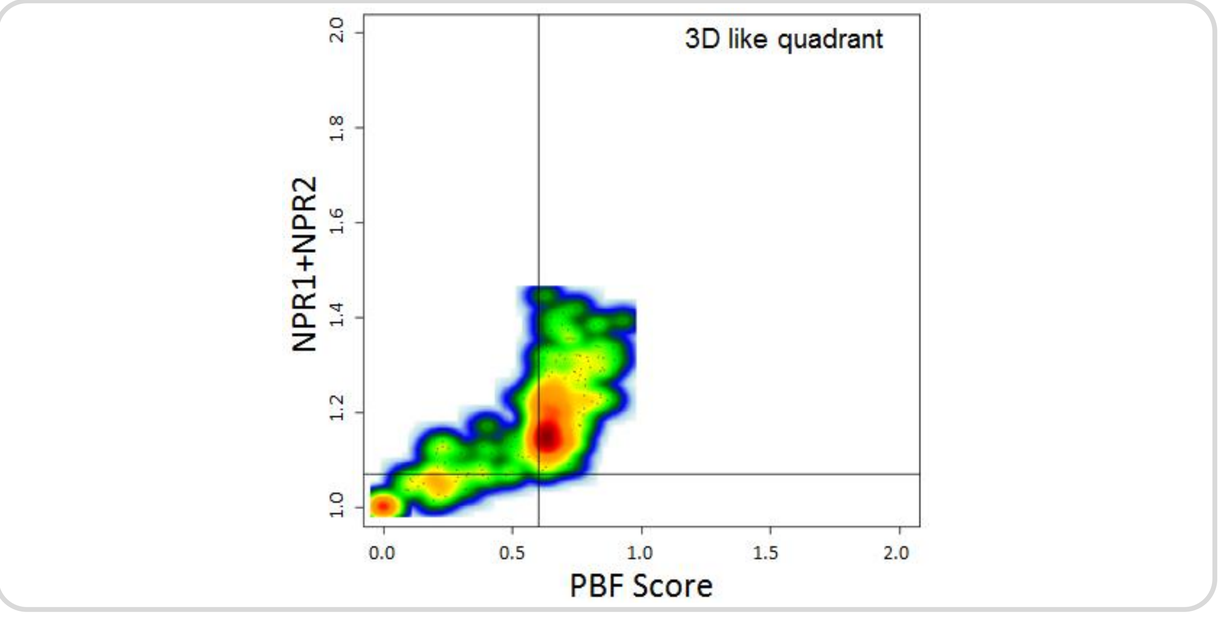

Figure 4: Density plot of PBF score vs the sum of the NPR1+2 for the 544,000 compounds taken from vendor libraries which passed the initial filtering criteria. The horizontal black line represents a cut-off for 3D molecules for NPR1+2 and the vertical line is the corresponding cut-off for PBF.

The Cocktail Scoring Function

In order to build diversity into the library a set of 2D and 3D descriptors were used to create a vector to compare compounds, these included the number of HBD and HBA, the number of rotatable bonds, the heavy atom count, the number of 5 and 6 membered rings, Fsp^3, molecular globularity and a 2D shape classifier derived from Murcko frameworks. The similarity of two molecules is calculated by finding the distance between the feature vectors of the two molecules. We aimed to have 90 cocktails of 4 compounds where 3 of the compounds were 3D-like. A multi-parameter optimiser was developed to enrich the diversity of the cocktails and of the entire library based on what we term the Cocktail Scoring Function. This is calculated by summing the dissimilarity measures, *d(i,j)*, of all pairs of compounds in a particular cocktail, *c*. The overall score, **S**, is calculated by adding the scores of all *n* cocktails.

$$S = \sum_{c=1,n} \sum_{i,j \in c} d(i,j)$$

The function is maximised through an iterative sampling of the 544,000 compound pool, exchanging single compounds and re-evaluating the score until the score does not change significantly. A generic algorithm This swapping is repeated 500 times per compound (*i.e.*, 2000 times per cocktail). After around 120 cycles across the full the 90 cocktail library.

The result of this optimisation routine was to build a 3D-enhanced library which was diverse both across the whole library and also within cocktails, thus allowing for good coverage of chemical space and rapid deconvolution of the cocktails. This library is now being screened at Evotec.

References

1) Michael R. Berthold, Nicolas Cebron, Fabian Dill et al., KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.
2) Mazanetz, M., Law, R. and Whittaker, M. 2013, Hit and Lead Identification from Fragments, in De novo Molecular Design (ed G. Schneider), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany
3) Congreve, M. Carr, R., Murray, C., Jhoti, H. (2003) A 'Rule of Three' for fragment-based lead discovery? Drug Discov. Today, 8(19), 876–877.
4) Sauer, W.H.B. and Schwarz, M.K. (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. J. Chem. Inf. Comput. Sci. 43, 987–1003
5) N Firth, N. Brown, J. Blagg, J. Chem. Inf. Model. 2012, 52, 2516−2525.